Human-Mouse Comparative Genomics: Successes and Failures to Reveal Functional Regions of the Human Genome

Len A. Pennacchio[1,2], Nadine Baroukh[1] and Edward M. Rubin[1,2]

[1]Genome Sciences Department, MS 84-171, One Cyclotron Road, Lawrence Berkeley National Laboratory, Berkeley, California, 94720, USA and [2]US Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, California, 94598, USA.

Short title: Human-Mouse Comparative Genomics

Correspondence and Proof Corrections:

Len A. Pennacchio

Genome Sciences Department

MS 84-171, One Cyclotron Road

Lawrence Berkeley National Laboratory

Berkeley, California, 94720, USA

Phone: 510-486-7498

Fax: 510-486-4229

Email: LAPennacchio@lbl.gov

Deciphering the genetic code embedded within the human genome remains a significant challenge despite the human genome consortium's recent success at defining its linear sequence (Lander et al. 2001; Venter et al. 2001). While useful strategies exist to identify a large percentage of protein encoding regions, efforts to accurately define functional sequences in the remaining ~97% of the genome lag. Our primary interest has been to utilize the evolutionary relationship and the universal nature of genomic sequence information in vertebrates to reveal functional elements in the human genome. This has been achieved through the combined use of vertebrate comparative genomics to pinpoint highly conserved sequences as candidates for biological activity and transgenic mouse studies to address the functionality of defined human DNA fragments. Accordingly, we describe strategies and insights into functional sequences in the human genome through the use of comparative genomics coupled with functional studies in the mouse.

Background

Mouse transgenesis experiments have constantly provided support for the universality of sequence-based regulatory information across vertebrates. Numerous examples exist where genes from a variety of vertebrates when introduced into mice as genomic transgenes express in a manner mimicking their expression in the natural host. One example of this is the human apolipoprotein A1 gene (*APOA1*) which has a well-described pattern of expression in the liver and intestines of both humans and mice. Indeed, in human *APOA1* transgenic mice, robust expression of the human gene in mouse

liver and intestines was observed, consistent with the mouse being able to recognize the regulatory sequences embedded within the human genomic transgene (Rubin et al. 1991). This *APOA1* study reflects data from a large number of mouse transgenesis experiments over the past 15 years which have repeatedly supported that despite the ~80 million years since the last common ancestor of humans and mice, regulatory information has been conserved and this supports the existence of a common gene regulatory vocabulary residing in the mammalian genome.

A particularly revealing mouse transgenesis study involved the generation and analysis of transgenic mice for a human gene for which there is no ortholog in the mouse genome. The human apolipoprotein (a) gene {*apo(a)*} recently arose in old-world monkeys and when a large human genomic transgene (250kb) containing *apo(a)* and flanking sequence was introduced into the mouse genome, its tissue expression pattern and components of its expression response to environmental factors mimicked that found in humans (Frazer et al. 1995). This study again highlights the existence of a highly conserved gene regulatory genetic code embedded in the noncoding sequence of mammals that determine neighboring gene expression characteristics.

Identification of a Gene Regulatory Element through Human-Mouse Comparative Genomics

One of the challenges following traditional mouse transgenesis experiments and the many reports of successful recapitulation of human gene expression in the mouse, is the downstream determination of the precise *cis*-regulatory sequences responsible for this activity. The recent availability of several vertebate genome sequences (human, mouse, rat, fugu, zebrafish (Aparicio et al. 2002; Dehal et al. 2002; Lander et al. 2001; Venter et al. 2001; Waterston et al. 2002) has allowed the exploitation of comparative sequence analysis to reveal conserved intervals in the human genome as candidates for explaining this biological activity (Duret et al. 1997; Hardison 2000; Hardison et al. 1997; Pennacchio et al. 2001b; Pennacchio et al. 2003). Since whole genome sequence datasets for human and mouse are the most advanced (Lander et al. 2001; Venter et al. 2001; Waterston et al. 2002), we will discuss the current power of comparing these two genomes as well as the potential limitations of this single pair-wise comparison.

As an example of how comparative genomics can be used as a starting point to lead biological experimentation, we previously compared human-mouse sequence in an approximately one megabase region (Mb) of human chromosome 5q31 (including five interleukins (IL) and 18 other genes) and its orthologous mouse region (Loots et al. 2000). This cross-species annotation of sequence identified 90 elements ≥100 bp that were conserved between human and mouse at a level of ≥70% identity (Figure 1A). Within this dataset, several previously characterized gene regulatory elements known to reside within this interval were readily identified by human-mouse sequence conservation, supporting the possibility of using such a strategy to identify gene regulatory elements.

To test the utility of comparative genomics to identify previously unknown gene regulatory elements, we studied the properties of a single conserved noncoding sequence (CNS1) located within the 15 kb interval between *IL-4* and *IL-13* (Figure 1B). This single element was chosen for detailed characterization based on its large size and high percent identity in the 1Mb interval (400 bp at approximately 87% identity between human and mouse). Furthermore, previous studies have suggested that *IL-4* and *IL-13* are co-regulated in T_h2 cells, raising the possibility that this single element might explain the co-regulation of these two genes. To characterize the function of CNS1, both transgenic and knockout mouse studies were performed (Loots et al. 2000; Mohrs et al. 2001). These independent *in vivo* strategies both revealed that CNS1 dramatically impacted on the expression of three human cytokine genes (*IL-4*, *IL-5*, and *IL-13*) separated by more than 120kb of sequence. For instance, in mice engineered to lack CNS1, a significant reduction in the number of T cells secreting IL-4 was found and this effect was not seen in Mast cells (Figure 1C) (Mohrs et al. 2001). Thus, conservation of sequence alone lead to the identification of a novel gene regulatory element which acts over long distances. Subsequent studies on CNS1 have further supported that this 400 bp element contains transcription factor binding sites which co-activate *IL-4*, *IL-5* and *IL-13* (Lee et al. 2001; Mohrs et al. 2001). It is interesting to note that while addional genes are found interspersed within these interleukin cluster, only the three interleukin genes appear to have altered expression when CNS1 was deleted *in vivo*.

This single study illustrated the complexity of long-range gene regulatory elements and the power of comparative biology to discover them. In the case of CNS1, as well as numerous other examples of previously identified gene enhancers, these elements are found within highly conserved human-mouse intervals which are devoid of flanking non-coding conservation, making their identification straightforward. These findings implied that the rapid scanning of the human genome for noncoding conservation with mouse should reveal a large number of human gene regulatory elements, but how well does this hold true?

Pitfalls: An Example where Comparative Genomics Fails to Reveal a Gene Regulatory Element

In the field of science, hypotheses are put forth and those which withstand rigorous testing are commonly reported as positive findings. Unfortunately, in addition to these positives stories, there are also numerous failed experiments that more often than not go unreported. While an increasing number of discoveries have been made using comparative genomics as a starting point with the hypothesis that *conserved sequences are fuctionally important*, failures have also occurred. One detailed example is provided below.

To identify gene regulatory elements within a four gene apolipoprotein cluster on human chromosome 11q23(Karathanasis 1985; Pennacchio et al. 2001a), we performed human-

mouse comparative analysis as a follow-up to the successful discovery of CNS1 within the interleukin gene cluster on human chromosome 5. Once again the goal was to find highly conserved human-mouse noncoding elements within this interval which could be tested for biological activity *in vivo*. Towards this goal, we were led to a ~600 bp human noncoding fragment that displayed ~70% identity with mouse which we chose to explore in further detail (Figure 2A). Similar to CNS1 in the interleukin cluster, this conserved sequence stood out discretely within a larger interval devoid of other noncoding conservation. This single finding supported that this human-mouse element has resisted "genetic drift", presumably due to functional constraints. The fact that it existed so prominently in a large interval containing four apolipoproteins with a complex expression pattern, and based on our previous experience with CNS1, suggested it too was a gene regulatory element.

To test this hypothesis, we engineered a bacterial artificial chromosome containing the entire human apolipoprotein gene cluster with loxP sites to flank the highly conserved sequence (Figure 2B). Our goal was to create two lines of transgenic mice; one that contained the human BAC plus the conserved element and a second that contained the human BAC minus the conserved element. This strategy was selected since it allowed us to compare the expression pattern of the human genes within the BAC in the two lines of transgenic mice in a position and copy number independent manner. This was achieved by breeding the original transgenics for the BAC plus the conserved element flanked by loxP sites with Cre-Recombinase expressing animals which produce a second line of mice where the conserved element was deleted (Figure 2B).

Examination of mice with the conserved element compared to mice lacking the element revealed no detectable difference in any of the neighboring apolipoproteins known expression pattern, despite extensive RNA analysis (Figure 2C). In addition, determination of these genes protein levels in plasma also indicated no differences despite the deletion of the conserved element. These studies indicate that under the *in vivo* conditions in which these elements were assessed, no function could be assigned to this conserved sequence. Whether it functions in gene regulation at another time point, environmental condition or performs non-gene regulatory roles remains unclear. Alternatively, the element could be functionless. As a second approach to test for gene regulatory properties, we fused this conserved sequence to a minimal reporter vector and generated transgenic mice. Again, these 600bp fragment was found to lack enhancer activity, in this case in 13.5 day embryos (data not shown). This example highlights the complexity of assigning function to highly conserved DNA element and the determination of what assays are the best to capture the endless number of functional possibilities. While human-mouse comparative genomics have provided the identity of numerous conserved elements with gene regulatory properties, many conserved elements are unlikely to be easily assigned a function. A key part of the interleukin CNS1 study was the detailed phenotypic analyses of *IL-4*, *IL-5* and *IL-13*. This particular phenotype was only found in stimulated T_h2 cells which were analyzed by flow-cytometry. Had a less sensitive phenotypic assay been performed, CNS1 would also appear non-functional. These two examples, the interleukin cluster on chromosome 5q31 and the apolipoprotein cluster on 11q23, provide early insights into the types of data expected to result from

comparative genomic driven studies. Having a strong understanding of a given genes complex expression pattern and phenotypic assays to assess this compexity is anticipated to greatly aid in the identification of gene regulatory elements.

Extrapolating Gene Regulatory Scans to the Whole Genome

The recent completion and comparison of a draft genome sequence of mouse to that of human revealed a striking amount of DNA conservation. In one study, it was found that ~40% of the human genome could be aligned to the mouse genome at the nucleotide level (Waterston et al. 2002). In a second study, separate analysis uncovered the identity of over one million discrete human-mouse conserved elements across the human genome (≥70% identity over ≥100bp) (Couronne et al. 2003). Further extrapolations from these studies strongly support the vast majority of human-mouse conservation is found in noncoding DNA. For instance, if 40% of the human genome can be aligned to mouse and yet only ~5% of the genome is found in mature mRNA transcripts, most human-mouse conservation can not be explained by this category of expressed DNA. In addition, of the greater than one million discrete human-mouse conserved elements, current estimates suggest that only ~200,000 of these elements are conserved as the result of exons. Thus again, current predictions suggest that a significant amount of conservation exists in noncoding human DNA (Couronne et al. 2003; Waterston et al. 2002). A key question that remains is what fraction of this noncoding DNA is functional and what biological processes do they perform? High-throughput strategies are currently

needed to categorize these large number of human-mouse conserved noncoding sequences.

One strategy to reduce the large number of human-mouse noncoding sequence elements for functional studies is to perform additional multi-species sequencing and analysis (Frazer et al. 2003; Mayor et al. 2000; Pennacchio et al. 2001b; Pennacchio et al. 2003; Schwartz et al. 2000). This can be achieved through the addition of a small number of more distantly-related species (such a fish, bird, or amphibian), or the use of a larger number of similarly distanced species (such as several additional mammals) (Bagheri-Fam et al. 2001; Cooper et al. 2003; Gilligan et al. 2002; Gottgens et al. 2002; Ureta-Vidal et al. 2003).

Deep Primate Sequence Comparisons to Reveal "Phylogenetic Shadows"

In contrast to distant cross-vertebrate sequence comparisons, a recently developed strategy for annotating genomes has been to perform deep sequence comparisons of evolutionary closely-related species (Boffelli et al. 2003). The general goal previously described for cross-species sequence comparisons is to use species of relatively distant phylogenetic positions to maximize the identification of functionally conserved sequences. However, this strategy fails in the search for species-specific genes and regulatory sequences such as those unique to primates. Recent comparison of the human and mouse genomes indicate that only ~80% of human-mouse genes have a 1:1

orthologous relationship (Waterston et al. 2002). Thus, there is a need for strategies to characterize the 20% of genes and regulatory elements that do not have a true ortholog in both humans and mice. For these studies, comparing human sequences to that of closer evolutionary species is warranted. Yet, the use of primate sequences for cross-species sequence comparisons is limited due to the high level of homology between these species.

"Phylogenetic shadowing" was developed to overcome the excessive sequence identity shared between two primates, making their use in cross-species sequence comparisons possible (Boffelli et al. 2003). The principle behind this strategy is to analyze orthologous sequence from numerous primate species to increase the sum of the evolutionary distance being compared. Rather than performing only pair-wise comparisons between human-mouse, "phylogenetic shadowing" compares a dozen or more different primate species. The additivity of these primate differences robustly defines regions of increased variation and "shadows" representing conserved segments (Figure 3A).

In its first use, "phylogenetic shadowing" proved successful in the identification of both exons and putative gene regulatory elements (Boffelli et al. 2003). This work generated and analyzed 13-17 primate species for several orthologous genomic segments. Examination of a single exon from four independent genes revealed highly conserved "shadows" which overlapped with these functionally important regions (one example is provided in Figure 3B for an exon of the apolipoprotein B gene). Further analysis of the

human apolipoprotein (a) gene (*apo(a)*) revealed highly conserved motifs embedded within the upstream promoter region. Functional characterization of these "phylogenetic shadows" compared to more variable flanking DNA supported their role in regulating *apo(a)* expression (Boffelli et al. 2003).

Additional analyses of these dataset suggest that less than a dozen primate sequence comparisons can suffice to detect functional sequences, provided they maximize the phylogenetic distance. In fact, in Figure 3B only 5 primates were examined and they proved successful in identifying a exon of the apolipoprotein B gene based on conservation. These species included human, talapoin, hanuman, spider monkey, and marmoset, which represent the most diverse primates within the large primate sequence dataset (Figure 3C). This initial success warrants further examination of this technique in other genomic intervals to determine its overall utility and suggest that this approach on a genome-wide scale will aid in the identification of both human exons and gene regulatory elements.

Conclusions

We have entered an era where the entire genomes of an increasing number of vertebrates have been sequenced and human-mouse whole genome comparisons are providing early insights into the realm of discovery possibilities. The single human-mouse pair-wise comparison has revealed new genes, regulatory elements, and a entire catalog of highly

conserved sequences with putative functionality. However, with this dataset, it has become clear that no single pair-wise comparison is suited to capture all biological activity. Current efforts to sequence a wide-variety of species, both evolutionarily closer and more distant from humans, are warranted (Boguski 2002; Sidow 2002). Clear examples exist of how human-mouse comparisons fail to capture known human funtional elements and support closer sequence comparisons. In addition, certain regions of the human and mouse genome are highly similar over long lengths thereby shielding the identification of highly conserved motifs for functional studies. Thus, the generation of a wide-ranging sequence dataset from a variety of vertebrates and beyond will provide useful information as to the genetic changes that have occurred over the evolutionary process and resulted in present day *Homo sapiens*.

Acknowledgements

This work was supported in part by the NIH-NHLBI Programs for Genomic Application Grant HL66681 and NIH Grant HL071954A through the U.S. Department of Energy under contract no. DE-AC03-76SF00098.

References

Aparicio S., Chapman J., Stupka E., Putnam N., Chia J.M., Dehal P., Christoffels A., Rash S., Hoon S., Smit A., Gelpke M.D., Roach J., Oh T., Ho I.Y., Wong M., Detter C., Verhoef F., Predki P., Tay A., Lucas S., Richardson P., Smith S.F., Clark M.S., Edwards Y.J., Doggett N., Zharkikh A., Tavtigian S.V., Pruss D., Barnstead M., Evans C., Baden H., Powell J., Glusman G., Rowen L., Hood L., Tan Y.H., Elgar G., Hawkins T., Venkatesh B., Rokhsar D., and Brenner S. 2002. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science* **297:** 1301-10.

Bagheri-Fam S., Ferraz C., Demaille J., Scherer G., and Pfeifer D. 2001. Comparative genomics of the SOX9 region in human and Fugu rubripes: conservation of short regulatory sequence elements within large intergenic regions. *Genomics* **78:** 73-82.

Boffelli D., McAuliffe J., Ovcharenko D., Lewis K.D., Ovcharenko I., Pachter L., and Rubin E.M. 2003. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299:** 1391-4.

Boguski M.S. 2002. Comparative genomics: the mouse that roared. *Nature* **420:** 515-6.

Cooper G.M., Brudno M., Green E.D., Batzoglou S., and Sidow A. 2003. Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes. *Genome Res* **13:** 813-20.

Couronne O., Poliakov A., Bray N., Ishkhanov T., Ryaboy D., Rubin E.M., Pachter L., and Dubchak I. 2003. Strategies and Tools for Whole-Genome Alignments. *Genome Res* **13:** 73-80.

Dehal P., Satou Y., Campbell R.K., Chapman J., Degnan B., De Tomaso A., Davidson B., Di Gregorio A., Gelpke M., Goodstein D.M., Harafuji N., Hastings K.E., Ho I., Hotta K., Huang W., Kawashima T., Lemaire P., Martinez D., Meinertzhagen I.A., Necula S., Nonaka M., Putnam N., Rash S., Saiga H., Satake M., Terry A., Yamada L., Wang H.G., Awazu S., Azumi K., Boore J., Branno M., Chin-Bow S., DeSantis R., Doyle S., Francino P., Keys D.N., Haga S., Hayashi H., Hino K., Imai K.S., Inaba K., Kano S., Kobayashi K., Kobayashi M., Lee B.I., Makabe K.W., Manohar C., Matassi G., Medina M., Mochizuki Y., Mount S., Morishita T., Miura S., Nakayama A., Nishizaka S., Nomoto H., Ohta F., Oishi K., Rigoutsos I., Sano M., Sasaki A., Sasakura Y., Shoguchi E., Shin I.T., Spagnuolo A., Stainier D., Suzuki M.M., Tassy O., Takatori N., Tokuoka M., Yagi K., Yoshizaki F., Wada S., Zhang C., Hyatt P.D., Larimer F., Detter C., Doggett N., Glavina T., Hawkins T., Richardson P., Lucas S., Kohara Y., Levine M., Satoh N., and Rokhsar D.S. 2002. The Draft Genome of Ciona intestinalis: Insights into Chordate and Vertebrate Origins. *Science* **298:** 2157-2167.

Duret L., and Bucher P. 1997. Searching for regulatory elements in human noncoding sequences. *Curr Opin Struct Biol* **7:** 399-406.

Frazer K.A., Elnitski L., Church D.M., Dubchak I., and Hardison R.C. 2003. Cross-species sequence comparisons: a review of methods and available resources. *Genome Res* **13:** 1-12.

Frazer K.A., Narla G., Zhang J.L., and Rubin E.M. 1995. The apolipoprotein(a) gene is regulated by sex hormones and acute-phase inducers in YAC transgenic mice. *Nat Genet* **9:** 424-31.

Gilligan P., Brenner S., and Venkatesh B. 2002. Fugu and human sequence comparison identifies novel human genes and conserved non-coding sequences. *Gene* **294:** 35.

Gottgens B., Barton L.M., Chapman M.A., Sinclair A.M., Knudsen B., Grafham D., Gilbert J.G., Rogers J., Bentley D.R., and Green A.R. 2002. Transcriptional regulation of the stem cell leukemia gene (SCL)--comparative analysis of five vertebrate SCL loci. *Genome Res* **12:** 749-59.

Hardison R.C. 2000. Conserved noncoding sequences are reliable guides to regulatory elements [In Process Citation]. *Trends Genet* **16:** 369-72.

Hardison R.C., Oeltjen J., and Miller W. 1997. Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. *Genome Res* **7:** 959-66.

Karathanasis S.K. 1985. Apolipoprotein multigene family: tandem organization of human apolipoprotein AI, CIII, and AIV genes. *Proc Natl Acad Sci U S A* **82:** 6374-8.

Lander E.S., Linton L.M., Birren B., Nusbaum C., Zody M.C., Baldwin J., Devon K., Dewar K., Doyle M., FitzHugh W., Funke R., Gage D., Harris K., Heaford A., Howland J., Kann L., Lehoczky J., LeVine R., McEwan P., McKernan K., Meldrim J., Mesirov J.P., Miranda C., Morris W., Naylor J., Raymond C., Rosetti M., Santos R., Sheridan A., Sougnez C., Stange-Thomann N., Stojanovic N., Subramanian A., Wyman D., Rogers J., Sulston J., Ainscough R., Beck S., Bentley D., Burton J., Clee C., Carter N., Coulson A., Deadman R., Deloukas P.,

Dunham A., Dunham I., Durbin R., French L., Grafham D., Gregory S., Hubbard T., Humphray S., Hunt A., Jones M., Lloyd C., McMurray A., Matthews L., Mercer S., Milne S., Mullikin J.C., Mungall A., Plumb R., Ross M., Shownkeen R., Sims S., Waterston R.H., Wilson R.K., Hillier L.W., McPherson J.D., Marra M.A., Mardis E.R., Fulton L.A., Chinwalla A.T., Pepin K.H., Gish W.R., Chissoe S.L., Wendl M.C., Delehaunty K.D., Miner T.L., Delehaunty A., Kramer J.B., Cook L.L., Fulton R.S., Johnson D.L., Minx P.J., Clifton S.W., Hawkins T., Branscomb E., Predki P., Richardson P., Wenning S., Slezak T., Doggett N., Cheng J.F., Olsen A., Lucas S., Elkin C., Uberbacher E., Frazier M., Gibbs R.A., Muzny D.M., Scherer S.E., Bouck J.B., Sodergren E.J., Worley K.C., Rives C.M., Gorrell J.H., Metzker M.L., Naylor S.L., Kucherlapati R.S., Nelson D.L., Weinstock G.M., Sakaki Y., Fujiyama A., Hattori M., Yada T., Toyoda A., Itoh T., Kawagoe C., Watanabe H., Totoki Y., Taylor T., Weissenbach J., Heilig R., Saurin W., Artiguenave F., Brottier P., Bruls T., Pelletier E., Robert C., Wincker P., Smith D.R., Doucette-Stamm L., Rubenfield M., Weinstock K., Lee H.M., Dubois J., Rosenthal A., Platzer M., Nyakatura G., Taudien S., Rump A., Yang H., Yu J., Wang J., Huang G., Gu J., Hood L., Rowen L., Madan A., Qin S., Davis R.W., Federspiel N.A., Abola A.P., Proctor M.J., Myers R.M., Schmutz J., Dickson M., Grimwood J., Cox D.R., Olson M.V., Kaul R., Shimizu N., Kawasaki K., Minoshima S., Evans G.A., Athanasiou M., Schultz R., Roe B.A., Chen F., Pan H., Ramser J., Lehrach H., Reinhardt R., McCombie W.R., de la Bastide M., Dedhia N., Blocker H., Hornischer K., Nordsiek G., Agarwala R., Aravind L., Bailey J.A., Bateman A., Batzoglou S., Birney E., Bork P., Brown

D.G., Burge C.B., Cerutti L., Chen H.C., Church D., Clamp M., Copley R.R., Doerks T., Eddy S.R., Eichler E.E., Furey T.S., Galagan J., Gilbert J.G., Harmon C., Hayashizaki Y., Haussler D., Hermjakob H., Hokamp K., Jang W., Johnson L.S., Jones T.A., Kasif S., Kaspryzk A., Kennedy S., Kent W.J., Kitts P., Koonin E.V., Korf I., Kulp D., Lancet D., Lowe T.M., McLysaght A., Mikkelsen T., Moran J.V., Mulder N., Pollara V.J., Ponting C.P., Schuler G., Schultz J., Slater G., Smit A.F., Stupka E., Szustakowski J., Thierry-Mieg D., Thierry-Mieg J., Wagner L., Wallis J., Wheeler R., Williams A., Wolf Y.I., Wolfe K.H., Yang S.P., Yeh R.F., Collins F., Guyer M.S., Peterson J., Felsenfeld A., Wetterstrand K.A., Patrinos A., Morgan M.J., Szustakowki J., de Jong P., Catanese J.J., Osoegawa K., Shizuya H., Choi S., and Chen Y.J. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

Lee G.R., Fields P.E., and Flavell R.A. 2001. Regulation of IL-4 gene expression by distal regulatory elements and GATA-3 at the chromatin level. *Immunity* **14:** 447-59.

Loots G.G., Locksley R.M., Blankespoor C.M., Wang Z.E., Miller W., Rubin E.M., and Frazer K.A. 2000. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288:** 136-40.

Mayor C., Brudno M., Schwartz J.R., Poliakov A., Rubin E.M., Frazer K.A., Pachter L.S., and Dubchak I. 2000. VISTA : visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* **16:** 1046-7.

Mohrs M., Blankespoor C.M., Wang Z.E., Loots G.G., Afzal V., Hadeiba H., Shinkai K., Rubin E.M., and Locksley R.M. 2001. Deletion of a coordinate regulator of type 2 cytokine expression in mice. *Nat Immunol* **2:** 842-7.

Pennacchio L.A., Olivier M., Hubacek J.A., Cohen J.C., Cox D.R., Fruchart J.C., Krauss R.M., and Rubin E.M. 2001a. An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing. *Science* **294:** 169-73.

Pennacchio L.A., and Rubin E.M. 2001b. Genomic strategies to identify mammalian regulatory sequences. *Nat Rev Genet* **2:** 100-9.

Pennacchio L.A., and Rubin E.M. 2003. Comparative genomic tools and databases: providing insights into the human genome. *J Clin Invest* **111:** 1099-106.

Rubin E.M., Ishida B.Y., Clift S.M., and Krauss R.M. 1991. Expression of human apolipoprotein A-I in transgenic mice results in reduced plasma levels of murine apolipoprotein A-I and the appearance of two new high density lipoprotein size subclasses. *Proc Natl Acad Sci U S A* **88:** 434-8.

Schwartz S., Zhang Z., Frazer K.A., Smit A., Riemer C., Bouck J., Gibbs R., Hardison R., and Miller W. 2000. PipMaker--a web server for aligning two genomic DNA sequences. *Genome Res* **10:** 577-86.

Sidow A. 2002. Sequence first. Ask questions later. *Cell* **111:** 13-6.

Ureta-Vidal A., Ettwiller L., and Birney E. 2003. Comparative genomics: genome-wide analysis in metazoan eukaryotes. *Nat Rev Genet* **4:** 251-62.

Venter J.C., Adams M.D., Myers E.W., Li P.W., Mural R.J., Sutton G.G., Smith H.O., Yandell M., Evans C.A., Holt R.A., Gocayne J.D., Amanatides P., Ballew R.M., Huson D.H., Wortman J.R., Zhang Q., Kodira C.D., Zheng X.H., Chen L.,

Skupski M., Subramanian G., Thomas P.D., Zhang J., Gabor Miklos G.L., Nelson C., Broder S., Clark A.G., Nadeau J., McKusick V.A., Zinder N., Levine A.J., Roberts R.J., Simon M., Slayman C., Hunkapiller M., Bolanos R., Delcher A., Dew I., Fasulo D., Flanigan M., Florea L., Halpern A., Hannenhalli S., Kravitz S., Levy S., Mobarry C., Reinert K., Remington K., Abu-Threideh J., Beasley E., Biddick K., Bonazzi V., Brandon R., Cargill M., Chandramouliswaran I., Charlab R., Chaturvedi K., Deng Z., Di Francesco V., Dunn P., Eilbeck K., Evangelista C., Gabrielian A.E., Gan W., Ge W., Gong F., Gu Z., Guan P., Heiman T.J., Higgins M.E., Ji R.R., Ke Z., Ketchum K.A., Lai Z., Lei Y., Li Z., Li J., Liang Y., Lin X., Lu F., Merkulov G.V., Milshina N., Moore H.M., Naik A.K., Narayan V.A., Neelam B., Nusskern D., Rusch D.B., Salzberg S., Shao W., Shue B., Sun J., Wang Z., Wang A., Wang X., Wang J., Wei M., Wides R., Xiao C., Yan C., Yao A., Ye J., Zhan M., Zhang W., Zhang H., Zhao Q., Zheng L., Zhong F., Zhong W., Zhu S., Zhao S., Gilbert D., Baumhueter S., Spier G., Carter C., Cravchik A., Woodage T., Ali F., An H., Awe A., Baldwin D., Baden H., Barnstead M., Barrow I., Beeson K., Busam D., Carver A., Center A., Cheng M.L., Curry L., Danaher S., Davenport L., Desilets R., Dietz S., Dodson K., Doup L., Ferriera S., Garg N., Gluecksmann A., Hart B., Haynes J., Haynes C., Heiner C., Hladun S., Hostin D., Houck J., Howland T., Ibegwam C., Johnson J., Kalush F., Kline L., Koduru S., Love A., Mann F., May D., McCawley S., McIntosh T., McMullen I., Moy M., Moy L., Murphy B., Nelson K., Pfannkoch C., Pratts E., Puri V., Qureshi H., Reardon M., Rodriguez R., Rogers Y.H., Romblad D., Ruhfel B., Scott R., Sitter C., Smallwood M., Stewart E., Strong R., Suh E.,

Thomas R., Tint N.N., Tse S., Vech C., Wang G., Wetter J., Williams S., Williams M., Windsor S., Winn-Deen E., Wolfe K., Zaveri J., Zaveri K., Abril J.F., Guigo R., Campbell M.J., Sjolander K.V., Karlak B., Kejariwal A., Mi H., Lazareva B., Hatton T., Narechania A., Diemer K., Muruganujan A., Guo N., Sato S., Bafna V., Istrail S., Lippert R., Schwartz R., Walenz B., Yooseph S., Allen D., Basu A., Baxendale J., Blick L., Caminha M., Carnes-Stine J., Caulk P., Chiang Y.H., Coyne M., Dahlke C., Mays A., Dombroski M., Donnelly M., Ely D., Esparham S., Fosler C., Gire H., Glanowski S., Glasser K., Glodek A., Gorokhov M., Graham K., Gropman B., Harris M., Heil J., Henderson S., Hoover J., Jennings D., Jordan C., Jordan J., Kasha J., Kagan L., Kraft C., Levitsky A., Lewis M., Liu X., Lopez J., Ma D., Majoros W., McDaniel J., Murphy S., Newman M., Nguyen T., Nguyen N., Nodell M., Pan S., Peck J., Peterson M., Rowe W., Sanders R., Scott J., Simpson M., Smith T., Sprague A., Stockwell T., Turner R., Venter E., Wang M., Wen M., Wu D., Wu M., Xia A., Zandieh A., and Zhu X. 2001. The sequence of the human genome. *Science* **291:** 1304-51.

Waterston R.H., Lindblad-Toh K., Birney E., Rogers J., Abril J.F., Agarwal P., Agarwala R., Ainscough R., Alexandersson M., An P., Antonarakis S.E., Attwood J., Baertsch R., Bailey J., Barlow K., Beck S., Berry E., Birren B., Bloom T., Bork P., Botcherby M., Bray N., Brent M.R., Brown D.G., Brown S.D., Bult C., Burton J., Butler J., Campbell R.D., Carninci P., Cawley S., Chiaromonte F., Chinwalla A.T., Church D.M., Clamp M., Clee C., Collins F.S., Cook L.L., Copley R.R., Coulson A., Couronne O., Cuff J., Curwen V., Cutts T., Daly M., David R., Davies J., Delehaunty K.D., Deri J., Dermitzakis E.T., Dewey C., Dickens N.J.,

Diekhans M., Dodge S., Dubchak I., Dunn D.M., Eddy S.R., Elnitski L., Emes R.D., Eswara P., Eyras E., Felsenfeld A., Fewell G.A., Flicek P., Foley K., Frankel W.N., Fulton L.A., Fulton R.S., Furey T.S., Gage D., Gibbs R.A., Glusman G., Gnerre S., Goldman N., Goodstadt L., Grafham D., Graves T.A., Green E.D., Gregory S., Guigo R., Guyer M., Hardison R.C., Haussler D., Hayashizaki Y., Hillier L.W., Hinrichs A., Hlavina W., Holzer T., Hsu F., Hua A., Hubbard T., Hunt A., Jackson I., Jaffe D.B., Johnson L.S., Jones M., Jones T.A., Joy A., Kamal M., Karlsson E.K., Karolchik D., Kasprzyk A., Kawai J., Keibler E., Kells C., Kent W.J., Kirby A., Kolbe D.L., Korf I., Kucherlapati R.S., Kulbokas E.J., Kulp D., Landers T., Leger J.P., Leonard S., Letunic I., Levine R., Li J., Li M., Lloyd C., Lucas S., Ma B., Maglott D.R., Mardis E.R., Matthews L., Mauceli E., Mayer J.H., McCarthy M., McCombie W.R., McLaren S., McLay K., McPherson J.D., Meldrim J., Meredith B., Mesirov J.P., Miller W., Miner T.L., Mongin E., Montgomery K.T., Morgan M., Mott R., Mullikin J.C., Muzny D.M., Nash W.E., Nelson J.O., Nhan M.N., Nicol R., Ning Z., Nusbaum C., O'Connor M.J., Okazaki Y., Oliver K., Overton-Larty E., Pachter L., Parra G., Pepin K.H., Peterson J., Pevzner P., Plumb R., Pohl C.S., Poliakov A., Ponce T.C., Ponting C.P., Potter S., Quail M., Reymond A., Roe B.A., Roskin K.M., Rubin E.M., Rust A.G., Santos R., Sapojnikov V., Schultz B., Schultz J., Schwartz M.S., Schwartz S., Scott C., Seaman S., Searle S., Sharpe T., Sheridan A., Shownkeen R., Sims S., Singer J.B., Slater G., Smit A., Smith D.R., Spencer B., Stabenau A., Stange-Thomann N., Sugnet C., Suyama M., Tesler G., Thompson J., Torrents D., Trevaskis E., Tromp J., Ucla C., Ureta-Vidal A., Vinson J.P., Von Niederhausern

A.C., Wade C.M., Wall M., Weber R.J., Weiss R.B., Wendl M.C., West A.P., Wetterstrand K., Wheeler R., Whelan S., Wierzbowski J., Willey D., Williams S., Wilson R.K., Winter E., Worley K.C., Wyman D., Yang S., Yang S.P., Zdobnov E.M., Zody M.C., and Lander E.S. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520-62.

Figure legends

Figure 1: Analysis of noncoding conservation on a region of human chromosome 5q31 containing a cluster of interleukin genes. **(A)** Distribution of 90 human-mouse conserved noncoding sequences in a one megabase region of human chromosome 5q31. These elements were selected based on the criteria of each displaying ≥70% identity over ≥100bp. Genes are indicated by vertical gray boxes with arrowheads to the left of the boxes depicting the orientation of transcription. To the right of the schematic horizontal arrows depict the positions of the conserved noncoding sequences with the most highly conserved 15 elements highlighted. **(B)** VISTA analysis showing a human-mouse genomic sequence comparison of the *IL-4* and *IL-13* region. 27kb of human sequence is depicted on the x-axis with gene annotation indicated above the plot. Exons are displayed as black rectangles and the gene orientation by the arrow's direction. Percent identity of the orthologous mouse sequence to human is plotted on the y-axis (50-100%). The graphical plot is based on sliding window analysis of the underlying genomic alignment, in this illustration a 100bp window is used which slides at 40bp nucleotide increments. The vertical arrow indicates the location of CNS1 which was identified based on it's high degree of conservation between human-mouse (VISTA peak). **(C)** Expression analysis of mice targeted for a deletion of CNS1. Mast and T cells were isolated from wildtype, herterozygous and homozygous mice for the deletion. In this example, T cells were stimulated with PMA and the number of IL-4 secreting cells was determined.

Figure 2: Identification and analysis of a highly conserved noncoding sequence in the *APOA1/C3/A4/A5* gene cluster. **(A)** A human-mouse VISTA plot displaying the level of genomic sequence conservation. In each panel 30 kbp of contiguous human sequence is depicted on the x-axis. Above each panel, horizontal arrows indicate known genes and their orientation with each exon depicted by a box (gene names are indicated above each arrow). The VISTA graphical plot displays the level of homology between human and the orthologous mouse sequence. Human sequence is represented horizontally and the percent similarity with the mouse sequence is plotted vertically (ranging from 50-100% identity). The vertical arrow indicates a highly conserved noncoding sequences. **(B)** Strategy for studying conserved noncoding sequences *in vivo*. A human BAC containing the apolipoprotein gene cluster is engineered to contain loxP sites flanking the conserved sequence of interest. Following the generation of a founder mouse, breeding experiments to Cre-recombinase expressing mice generate a second line of animals with deletions for the conserved element of interest. **(C)** RNA analysis of trangenic mice (Tg) containing the conserved element (CNS) compared to transgenic mice lacking the element (ΔCNS). A wildtype control mouse is also provided (CT). Liver and intestine total RNA were prepared and hybridized with human-specific probes for *APOA1*, *APOC3*, *APOA4* and *APOA5*. Mouse beta-actin was used as an internal control. No differences were detected in transcript levels from transgenic animals containing the conserved element compared to animals lacking it.

Figure 3: "Phylogenetic shadowing" of closely-related species. **A)** The alignment and comparison of sequences from multiple primate species reveal sequences which have

been conserved across most species, making them candidates for being functionally relevant due to presumed evolutionary constraint at these sites. **B)** Primate-specific "phylogenetic shadowing" reveals a previously defined exon for the apolipoprotein B gene (*APOB*). On the x-axis a variation score is provided with more negative scores indicating less variable regions and on the y-axis 1500 base pairs of human sequence is displayed. The known *APOB* exon in this interval is depicted by a solid black line within the plot. Note the decreased amount of primate variation in regions corresponding to the exon. **C)** Primate phylogenetic tree based on a single genomic interval. A carefully selected set of species which maximize phylogenetic distance (boxed) can capture the majority of the "phylogenetic shadows" and thus can reduce the amount of genomic sequence information required.